Session: 33 Abstract Number: 5259
Xiaoqiang.wang@rennes.inra.fr

The repercussions of statistical properties of interval mapping methods on eQTL detection

Xiaoqiang Wang[1], Jean-Michel Elsen[2], Hélène Gilbert[3], Carole Moreno[2], Olivier Filangi[1], Pascale Le Roy[1]

[1]UMR598, Rennes, [2]UR631 Toulouse, [3]UMR337 Jouy en Josas, INRA,France

Summary

QTL detection on a huge amount of phenotypes, like eQTL detection on transcriptomic data, highlights the statistical properties of interval mapping methods. One of the steadiest outcomes is the high number of eQTL detected on markers locations. The aim of this communication is to describe QTL detection in this particular context through the use of simulated data. Designs of sib families were simulated and analyzed using the QTLMAP software. Different parameters, such as the population size, the QTL effect, the QTL location, the number of markers or the density of the genetic map, were taken into account. Simulations under the no QTL hypothesis showed that, whatever the location, i.e. on a marker or between two markers, the nominal test statistics follows a χ^2 distribution with a number of degrees of freedom depending on the number of parents. Simulations under the one QTL hypothesis confirmed that the estimated location of the QTL is biased. Indeed, it is closer to markers locations than it should be, which is even more noticeable towards the bounds of the linkage group. The lower the QTL effect, the higher this bias. The repercussions of the above on eQTL detection are discussed.

These results are obtained through the EC-funded FP6 Project "SABRE".

1 Introduction

Recently, we analyzed high throughput phenotypes, like transcriptomic data, from several experimental designs in pig and poultry. We used interval mapping procedures and we considered each transcript as one trait in a trait by trait analysis. We observed that the number of eQTL detected on markers locations is higher than between markers. This property of QTL detection methods was known (Walling *et al.*, 2001) but is highlighted in this particular context because of the huge amount of analyzed phenotypes. The aim of this communication is to describe this bias on the estimation of QTL location through the use of simulated data. Designs of sib families were simulated and analyzed using the QTLMAP software.

2 Simulations under H0

Under the null hypothesis of no QTL, it would be expected that the probability of the estimated QTL locations across the chromosome should be identical, in other words, the estimated QTL location should be uniformly distributed on the chromosome. But according to many simulations in such a case, we found that the probability for the estimated QTL location at markers loci is higher when compare to non-marker location. We observed this bias through an experimental design in pig. There are 16 markers, located respectively at 0cM, 15cM, 25cM, 63cM, 77cM, 88cM, 95cM, 104cM, 116cM, 122cM, 126cM, 143cM, 161cM, 195cM, 205cM, 213cM, on the chromosome SSC1. After having performed 2000 simulations in a population of 4 sires and 325 offspring, an empirical distribution of the estimated QTL location is given in the figure 2.1, which clearly shows that under the null hypothesis of no QTLs, a large proportion of QTLs are estimated to be at a marker position. In order

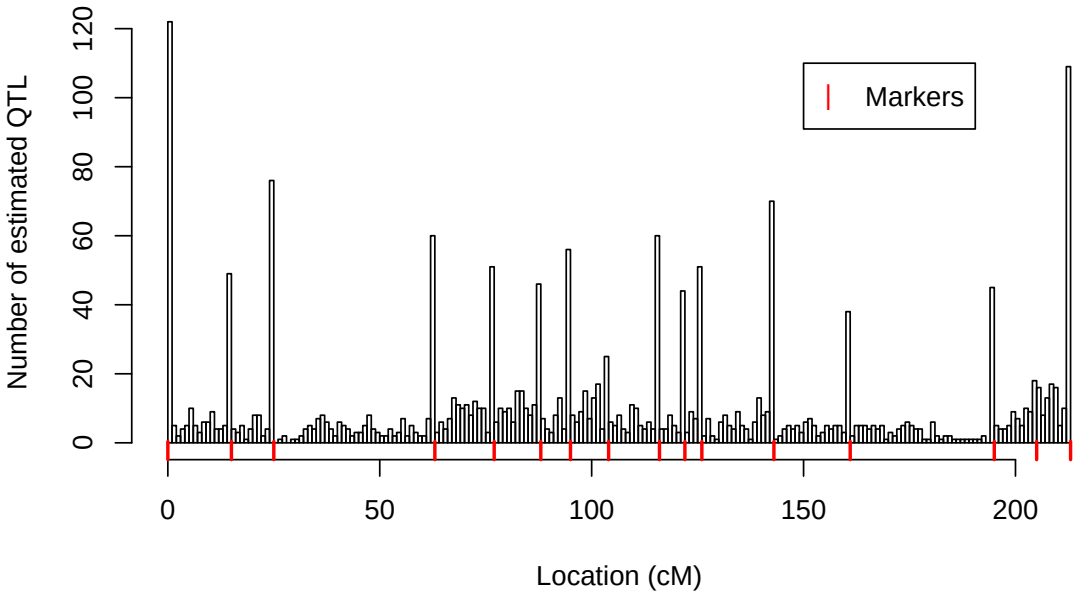

Figure 2.1: Empirical distribution of the estimated QTL location across a chromosome of 213cM. Results are based on 2000 simulations of population of 4 sires and 325 offspring, using 13 markers.

to study the element that give rise to them, we checked if the significance threshold changes among locations in the chromosome or not, in other words, does significance threshold depends on the marker or non-marker location? Significance thresholds in QTL analysis are often calculated by studying the distribution of the test statistics in the case of null hypothesis of no QTL by simulation. The histograms of figure 2.2 show the empirical distribution of some locations in the chromosome (both markers and non-markers). The LRT score at each location follows a χ^2 distribution with the degrees of freedom in [4.05,4.55] by the Kolmogorov-Smirnov test using a $\alpha = 0.01$. If the level of the test statistics depends on the location, i.e. at markers or between

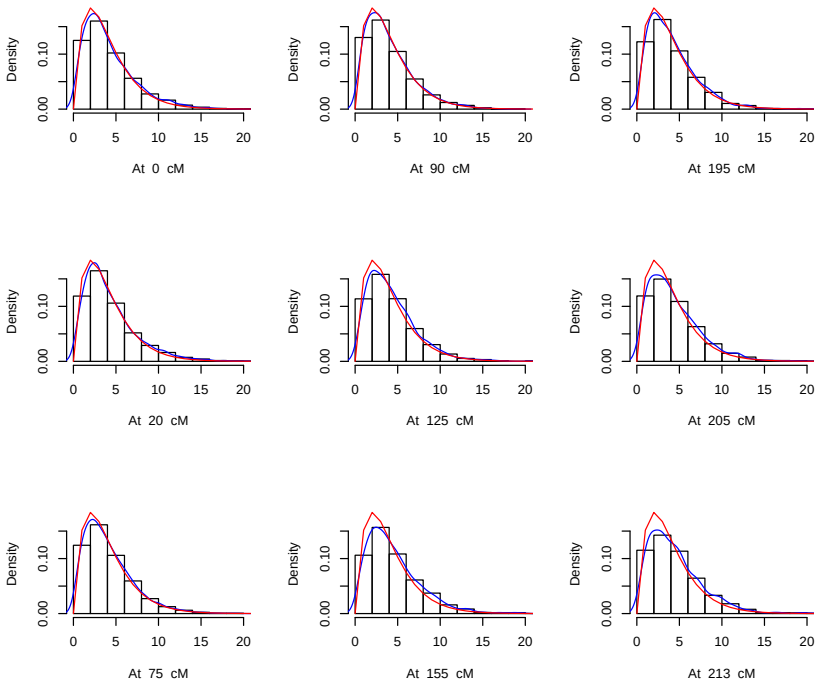

Figure 2.2: Empirical distribution of LRT score,the red line represent the density of χ^2 with 4 d.F., the blue line represent the density obtained by Kernel method.

markers, the degrees of freedom at markers locations should be greater than at non-markers locations. The figure 2.3 shows the degree of freedom of χ^2 distribution at each location. We noted that the distribution of LRT at markers was not obviously different than between markers.

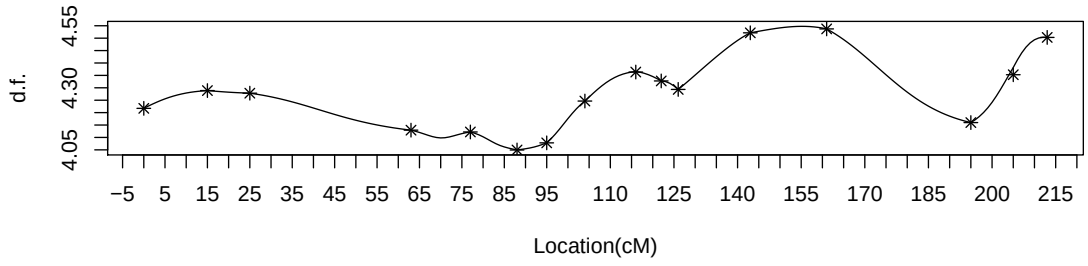

Figure 2.3: The degrees of freedom of χ^2 distribution at each location, the stars are markers.

Simulations under the no QTL hypothesis showed that, whatever the location, ie on a marker or between two markers, the nominal test statistics follows a χ^2 distribution. Theoretically, for both marker and non-marker locations, the LRT asymptotically follows a central χ^2 distribution with degree of freedom depending on the number of parameters fixed under H0, here the number of sires. Hence as the number of progeny increases, the distribution of LRT will be getting closer to

the same χ^2 distribution. In conclusion, we verified that the bias observed on the estimation of the QTL location is not due to the significance threshold.

3 Simulations under H1

In the case of having a QTL, different parameters, such as the population size, QTL effect, markers density and QTL location were taken into account. We focused on the bias caused by these parameters in this study.

We first simulated a QTL at 10cM and 3 markers at 0cM, 20cM, 40cM. The QTL effect always is 1σ. To understand the bias influenced by the population size, we changed the number of individual between 100 (5 sires, 1 dam per sire and 20 offspring per dam), 300 (5 sires, 2 dams per sire and 30 offspring per dam) and 800 (5 sires, 4 dams per sire and 40 offspring per dam). The empirical distribution of the estimated QTL location is shown in figure 3.1 and the proportion of estimated QTL co-localized with a marker is shown in table 1.

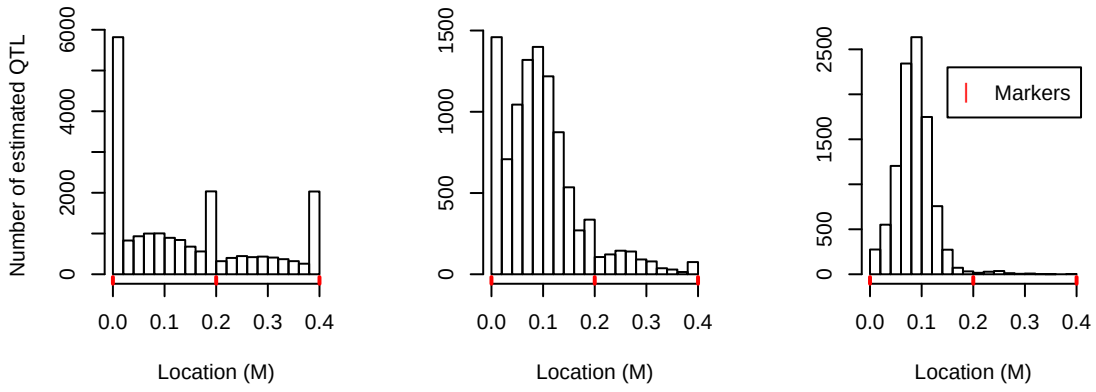

Figure 3.1: Empirical distributions of the estimated QTL location in a linkage group of 40cM. Results are based on 20000 simulations for first graph, 10000 simulations for the others.

Table 1: Proportion of estimated QTL co-localized with one marker

$n = 100$	$n = 300$	$n = 800$
0.44	0.113	0.013

It can be seen that the proportion of estimated QTL co-localized with marker decreases when the population size increases. The estimated looks like getting more and more exact as number of progeny increases.

To check the bias extent depending on the markers density, we simulated a QTL at 25cM in a population of 5 sires, 1 dam per sire, and 20 offspring per dam. The QTL effect is always fixed on 1σ and we increased the number of markers in a linkage group of 60cM between 3, 4 and 7. Then the mean square error (MSE) of estimated

QTL location is used to evaluate this bias. Results are given in figure 3.2 and in table 2.

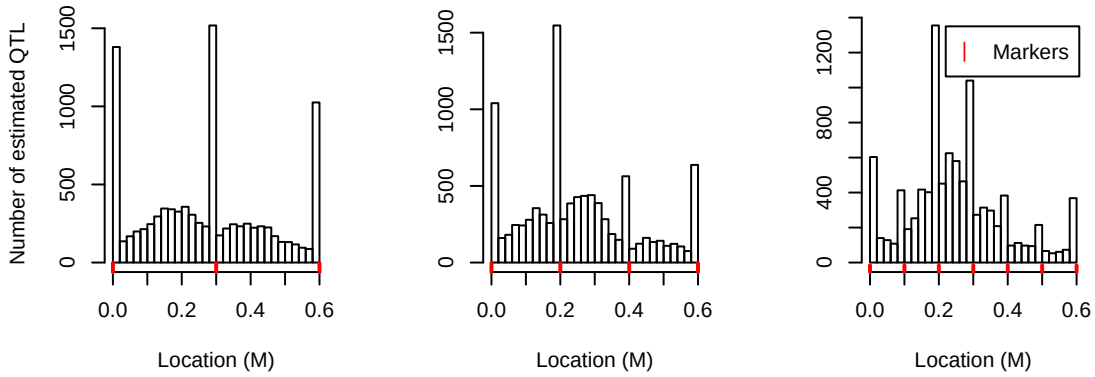

Figure 3.2: Empirical distributions of the estimated QTL location in a linkage group of 60cM. Results are based on 10000 simulations.

Table 2: Proportion of estimated QTL co-localized with marker and MSE of estimated QTL location

	3 markers	4 markers	7 markers
Proportion	0.368	0.348	0.379
MSE	0.0348	0.0277	0.0197

According to MSE, we found that, when the density increases, even if the proportion of estimated QTL locations at marker location is more or less the same, the estimated location looks like getting more and more exact.

It is possible that the true QTL location could affect the proportion of estimated QTL location at marker location. Hence, we performed 5000 simulations in a population of 5 sires, 1 dam per sire and 20 offspring per dam when the QTL is respectively fixed at 10cM, 30cM, 50cM in a linkage group of 60cM. And there are 4 markers at 0cM, 20cM, 40cM and 60cM respectively. The QTL effect always is 1σ. Table 3 shows a weak influence of the true QTL location for the bias. It seems that the estimated QTL location is more exact when the true QTL is located in the middle of the whole linkage group than when the true QTL is located at the other sites.

Table 3: Proportion of estimated QTL co-localized with marker and MSE of estimated QTL location

	QTL at 10cM	QTL at 30cM	QTL at 50cM
Proportion	0.42	0.34	0.40
MSE	0.0398	0.0304	0.0372

To observe how the QTL effect affect the estimation of QTL location, a QTL is simulated at 10cM in a population of 5 sires, 1 dam per sire, and 20 offspring per dam. There are 3 markers at 0cM, 20cM and 40cM. The QTL effect varies between 0.5σ, 1σ and 2σ. The empirical distribution of the estimated QTL location is shown in figure 3.3 and the proportion of estimated QTL co-localized with a marker is shown in table 4.

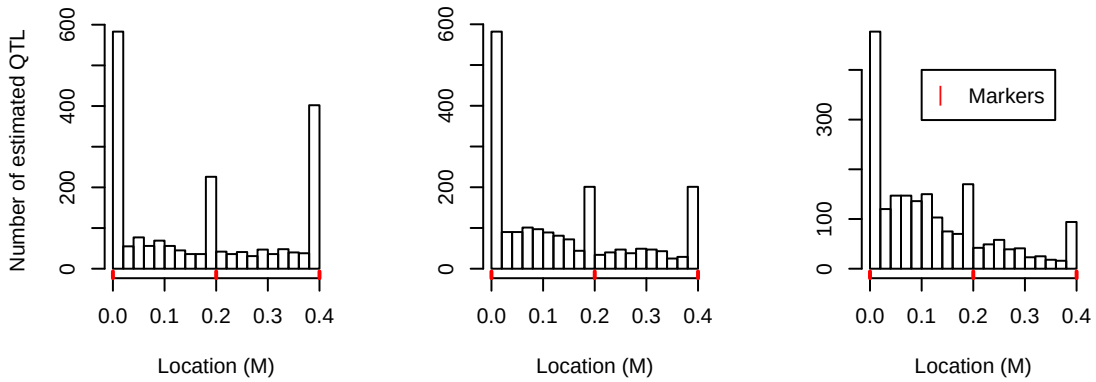

Figure 3.3: Empirical distributions of the estimated QTL location in a linkage group of 40cM. Results are based on 20000 simulations for first graph, 10000 simulations for the others.

Table 4: Proportion of estimated QTL co-localized with one marker

0.5σ	1σ	2σ
0.588	0.435	0.317

Results shown in figure 3.3 and the table 4 indicate a small QTL effect highly affects the bias.

4 Discussion

Under the null hypothesis of no QTL, we noticed that the proportion of the estimated location of putatif QTL at markers locations was higher in comparison with non-marker locations. However, the significance threshold is not a reason to give rise to them.

In the case of having a QTL, simulations confirmed that some parameters play a role affecting the bias of the estimated QTL location. Results display when the population size increases or when the markers density increases or when the QTL effect increases or when the real QTL location tends to the middle of linkage group, the estimated QTL location will be getting more and more exact.

Let $h_j(p)$ denote the probability of haplotypes transmission for individual conditional on the genotypes and the markers locations for location p. Let y_i represent

the phenotype of individual j. The calculation of the linearized LRT can be written as

$$\mathrm{LRT}^p = \left[\sum_j y_j \cdot \frac{[1 - 2h_j(p)]}{\sqrt{\sum_j [1 - 2h_j(p)]^2}} \right]^2$$

The estimated QTL effect is calculated by equation (1)

$$\hat{a} = \frac{\sum_j y_j (1 - 2h_j(p))}{\sum_j (1 - 2h_j(p))^2} \tag{1}$$

Equation (1) highlights how the marker location affects the calculation of the QTL effect. When the markers informativity is not complete, the probability of halpotypes transmission varies between 0 and 1 and, consequently, there is a possible confusion between the two estimated parameters: the QTL effect and the QTL location.

Serval authors (e.g. Spelman *et al.*, 1996; Walling *et al.*,1998) noted that a large proportion of the estimated QTL co-localized with marker, and questioned whether there is a bias in the regression method. Walling *et al.* (2001) predicted the expected proportion of QTLs with locations estimated to be at the location of a marker through studying the statistical property of the regression coefficients. Under the null hypothesis of no QTL between 2 flanking markers, Walling *et al.*(2001) predicted the proportion of putative QTLs placed at the flanking markers is

$$0.5 + \frac{\arcsin(1 - 2r)}{\pi}$$

where r denote the recombination rate between the flanking markers. Results look very similar to the ones we came up with by the interval mapping approach as shown in table 5.

Table 5: Comparison of the bias between the regression method and the maximum likelihood

Distance between 2 markers	10cM	20cM	40cM	100cM
Interval Mapping (QTLMAP)	82.6%	74.4%	66.6%	55.0%
Linear regression	80.5%	73.4%	64.8%	54.3%

Previous results are under the condition of independence between phenotypes. We further must use transcriptome data in the case of multivariate analysis. In this case, the bias may be reduced or disappear. In the maximum likelihood for QTL detection, Gilbert *et al.* (2003) confirmed the advantage of the multivariate analysis in comparison with univariate.

Acknowledgements

These results are part of the SABRE research project that has been co-financed by the European Commission, within the 6th Framework Programme, contract No. FOOD-CT-2006-016250.

References

[1] Gibert H, Le Roy P. (2003) Multidimensionnalité pour la détection de génes influençant des caractéres quantitatifs Application á l'espéce porcine. *PhD thesis. Institut National Agronomique Paris-Grignon, 144p.*

[2] Haley, C.S. & Knott, S.A. (1992) A simple regression method for mapping quantitative trait loci in line crosses using flanking markers. *Herdity 69, 315-324.*

[3] Lander, E.S. & Botstein, D.(1989). Mapping Mendelian factors underlying quantitative traits using RFLP linkage maps *Genetics 121, 185-199.*

[4] Spelman, R.J., Coppieters, W., Karim, L., van Arendonk, J.A.M. & Bovenhuis, H.(1996). Quantitative trait loci for five milk production traits on chromosome six int Dutch Holstein-Friesian population. *Genetics 144, 1799-1808.*

[5] Walling, G.A., Visscher, P.M. & Haley, C.S. (1998) A comparison of bootstrap methods to construct confidence intervals in QTL mapping. *Genetical Research 71, 171-180.*

[6] Walling G.A., Haley C.S., Perez-Enciso M., Thompson R., Visscher P.M. (2001) On the mapping of quantitative trait loci at marker and non-marker locations. *Genetical Research 79 , 97-106.*